21.2: CLT-based Inference for Difference in Means

Stat 120 | Fall 2025

Prof Amanda Luby

1 One-sample CLT inference in R

When we are performing a t-test or computing a t-based confidence interval for a mean, we can use the `t.test()` function in R. Here is a template:

```
t.test(y ~ 1, data = dataset, mu = null, direction = "two.sided")
```

where

- `y` is the column name for the variable
- `dataset` is the name of the data set
- `null` is the value of μ specified in the null hypothesis
- `alternative` specifies which tail corresponds to the p-value. The options are `"two.sided"`, `"less"`, and `"greater"`.
- `conf.level` can be added to change the confidence level of the interval returned

1. The code below reads in the `FloridaLakes` dataset from class last week. Use the `t.test()` function to perform the CLT-based t-test to see whether the average `MaxMercury` is less than 1 or not. Your results should match the "by hand" results from class. (Slides here).

```
lakes <- read.csv("http://www.lock5stat.com/datasets1e/FloridaLakes.csv")
```

2 Two-sample CLT inference in R

When we are comparing the mean of two independent groups, we again use the `t.test()` function in R. Here is a template:

```
t.test(y ~ x, data = dataset, mu = null, direction = "two.sided", paired = FALSE)
```

where

- `y` is the column name for the response variable
- `x` is the column name for the explanatory (grouping) variable
- `dataset` is the name of the data set
- `null` is the value of $\mu_1 - \mu_2$ specified in the null hypothesis
- `alternative` specifies which tail corresponds to the p-value. The options are `"two.sided"`, `"less"`, and `"greater"`.
- `conf.level` can be added to change the confidence level of the interval returned
- if `x` and `y` represent paired measurements, change `paired` to `TRUE` to perform a matched-pairs t-test

1. Researchers Holdgate et al. (2016) studied walking behavior of elephants in North American zoos to see whether there is a difference in average distance traveled by African and Asian elephants in captivity. They put GPS loggers on 33 African elephants and 23 Asian elephants, and measured the distance (in kilometers) the elephants walked per day.

```
elephants <- read.delim("http://www.rossmanchance.com/iscam3/data/Elephants.txt")
```

(a) Using EDA, compare the distribution of distance for each species. What do you learn?

(b) Determine the sample sizes, sample means, and the sample standard deviations for each group. Some starter code is below (you will need to remove the line that says `eval = FALSE`)

```
elephants %>%
  group_by(<var_name>) %>%
  summarize(
    mean = mean(<var_name>),
    sd = sd(<var_name>),
    n = n()
  )
```

(c) Are the conditions met for the t-procedures to be valid for these data?

(d) Use R to construct a 90% confidence interval for the difference between the average walking distance of the African and Asian elephant populations. Include one-sentence interpretation of your interval. (When interpreting a confidence interval for the difference, you should always clarify the direction of subtraction and which population parameter is larger.)